Evaluating Topic Quality with Posterior Variability (PV)

Linzi Xing[1]; Michael J. Paul[2]; Giuseppe Carenini[1]
[1]University of British Columbia, [2]University of Colorado Boulder
{lzxing, carenini}@cs.ubc.ca, mpaul@colorado.edu

University of Colorado Boulder

New Metric: Topic Posterior Variability (PV)

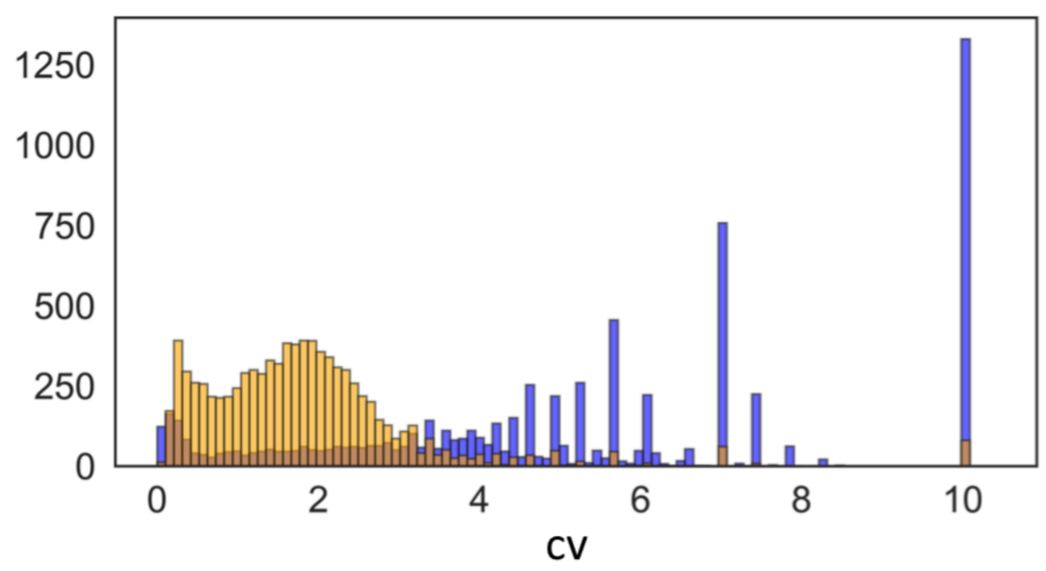
Topic posterior variability (PV) measures the degree of a topic's corpus-wide variability during Gibbs sampling, a posterior inference algorithm.

Standard deviation of estimates / Mean of estimates

$$cv_{dk} = \sigma_{dk} / \mu_{dk}$$

$$PV(k) = std(cv_{1k}, cv_{2k}, \cdots, cv_{Dk})$$

Example: The *cv* distributions of two example topics across NYT corpus. Human rating for topic a and b are 3.4 and 1 respectively.



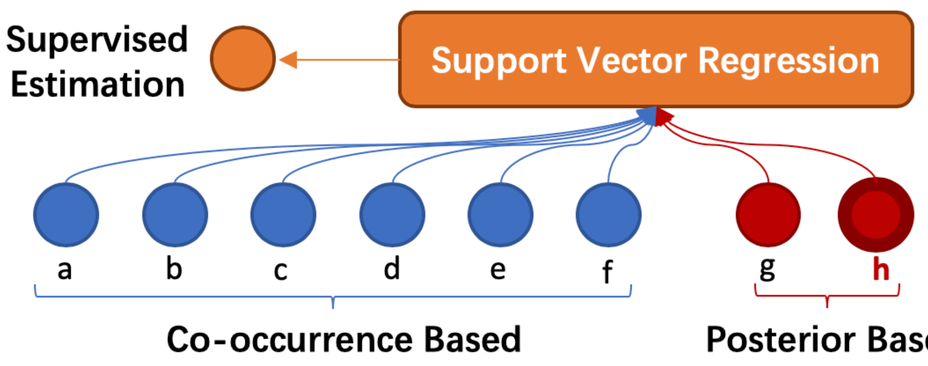
Topic a: financial, banks, bank, money, debt, fund, loans, investors, funds, hedge

Topic b: world, one, like, good, even, know, think, get, many, got

Existing Topic Quality Metrics + New One

- **Co-occurrence Based**
  - a. CV [1]
  - b. CP [1]
  - c. Distributional Similarity (DS) of Word Pairs [2]
  - d. Normalized Pointwise Mutual Information (NPMI) [3]
  - e. Pointwise Mutual Information (PMI) [4]
  - f. Topic Coherence [5]
- **Posterior Based**
  - g. Topic Stability [6]
  - **h. Topic Posterior Variability (PV) (new metric)**

Topic Quality Estimator



Conclusions

- Our proposed **topic posterior variability (PV)** is more accurate than previous methods when tested against human topic quality judgment.
- A **supervised topic quality estimator** delivers even better results by assembling multiple metrics.

Datasets

- **20NG** :
  9,347 paragraphs categorized into 20 classes.
- **Wiki :**
  10,773 Wikipedia articles written in simple English.
- **NYT :**
  8,764 New York Times articles from April to July, 2016.

❖ 100 topics for each dataset.

❖ The gold-standard annotation for the quality of each topic is the mean of 4-scale human ratings from five annotators.

Variability vs Earlier Topic Quality Metrics

Method	20NG	Wiki	NYT	Mean
CV[1]	.129	.385	.248	.254
CP[1]	.378	.403	.061	.280
DS[2]	.461	.423	.365	.416
NPMI[3]	.632	.568	.639	.615
PMI[4]	.602	.550	.623	.591
Coherence[5]	.280	.102	.535	.305
Stability[6]	.230	.137	.322	.230
Variability	**.679**	**.703**	**.774**	**.719**

The Pearson's r correlation with human judgments for topic posterior variability and earlier existing topic quality metrics.

Estimator vs Variability

Test	Train			Mean	Variability
20NG	Wiki	NYT	Wiki+NYT	**.801**	.679
	.790	.804	.810		
Wiki	20NG	NYT	20NG+NYT	**.716**	.703
	.707	.731	.710		
NYT	20NG	Wiki	20NG+Wiki	.770	**.774**
	.762	.775	.773		

The comparison of Pearson's r correlation with human rating between the topic posterior variability and the topic quality estimator.

Ablation Study For Topic Quality Estimator

References

[1] Roder, M.; Both, A.; Hinneburg, A. 2015. Exploring the space of topic coherence measures. In Proc. of WSDM 2015.

[2] Aletras, N.; Stevenson, M. 2013. Evaluating topic coherence using distributional semantics. In Proc. of IWCS 2013.

[3] Lau, J. H.; Newman, D.; Baldwin, T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proc. of EACL 2014.

[4] Newman D.; Lau, J. H.; Grieser, K.; Baldwin, T. 2010. Automatic evaluation of topic coherence. In Proc. of NAACL 2010.

[5] Mimno, D.; Blei,D. 2011. Bayesian checking for topic models. In Proc. of EMNLP 2011.

[6] Xing, L.; Paul, M. J. 2018. Diagnosing and improving topic models by analyzing posterior variability. In Proc. of AAAI-18.

UBCNLP Group

Original Paper